Extracting Keyphrases from Research Papers using Citation Networks

Cornelia Caragea, Joint Work with Sujatha Das Gollapalli
Computer Science and Engineering, University of North Texas

Why Keyphrase Extraction?

- Large number of scholarly documents on the Web.

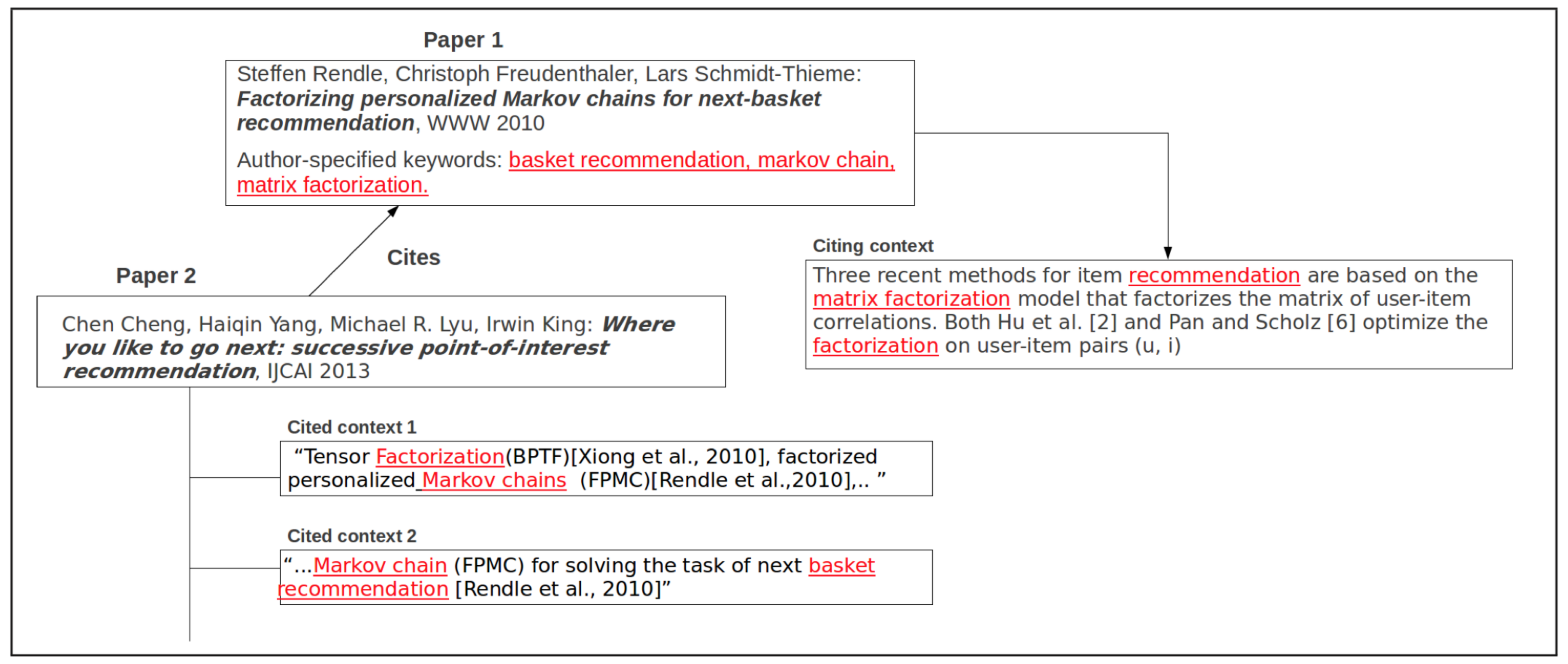
- The most important parts or "concepts" in these documents are not always directly available, but need to be gleaned from the multitude of details in documents.
- During these "big data" times, keyphrases associated with research papers can allow for *efficient processing of more information in less time*.
  - Useful in many ML and IR applications such as topic tracking, information filtering, and search.
- Keyphrase extraction is defined as the problem of automatically extracting descriptive phrases or concepts from a document.

Previous Approaches to Keyphrase Extraction

- Most existing keyphrase extraction techniques used only the textual content of the target document [Mihalcea & Tarau(2004), Liu et al.(2010)Liu, Huang, Zheng, & Sun].
- Wan and Xiao [Wan & Xiao(2008)] addressed this simplification using a model that incorporates a local neighborhood of a document for extracting keyphrases.
  - However, their neighborhood is limited to textually-similar documents.
- We posit that, in addition to a document's textual content and textually-similar neighbors, other informative neighborhoods exist in research document collections that have the potential to improve keyphrase extraction.

From Data to Knowledge

- Scientific research papers typically propose new problems or extend the state-of-the-art for existing research problems.
  - It is common to find in a document, relevant, previously-published research papers cited in appropriate *contexts*.
  - Such citations between research papers give rise to a large network of interlinked documents, commonly referred to as the *citation network*.
- In a citation network, information flows from one paper to another via the citation relation [Shi et al.(2010)Shi, Leskovec, & McFarland].
  - This information flow as well as the influence of one paper on another are specifically captured by means of *citation contexts* (i.e., short text segments surrounding a paper's mention).
  - These contexts are not arbitrary, but they serve as "micro summaries" of a cited paper.



- *Can citation networks improve the performance of keyphrase extraction? Since citation contexts capture how papers influence each other along various aspects, e.g., topicality, domain of study, and algorithms, how can we use these "micro summaries" in keyphrase extraction models?*

Proposed Approach: CiteTextRank

We propose CiteTextRank, a fully unsupervised graph-based algorithm that incorporates evidence from multiple sources (citation contexts as well as document content) in a flexible manner to extract keyphrases.

General steps for algorithms for unsupervised keyphrase extraction:

1. Extract candidate words or lexical units from the textual content of the target document by applying stopword and parts-of-speech filters.
2. Score candidate words based on some criterion.
   - For example, in the TFIDF scoring scheme, a candidate word score is the product of its frequency in the document and its inverse document frequency in the collection.
3. Finally, score consecutive words, phrases or n-grams using the sum of scores of individual words that comprise the phrase [Wan & Xiao(2008)]. Output the top-scoring phrases as predictions.

CiteTextRank incorporates information from *citation contexts* while scoring candidate words in step 2.

Graph Construction in CiteTextRank

Let d be the target document and $\mathcal{C}$ be a citation network such that $d \in \mathcal{C}$.

- Definitions:
  - A *cited context* for d is defined as a context in which d *is cited* by some paper d_i in the network.
  - A *citing context* for d is defined as a context in which d *is citing* some paper d_j in the network.
  - The content of d comprises its *global context*.
- Let T represent the types of available contexts for d,i.e., the *global* context of d, $\mathcal{N}_d^{Ctd}$, the set of *cited* contexts for d, and $\mathcal{N}_d^{Ctg}$, the set of *citing* contexts for d.

We construct an undirected graph, $G = (V, E)$ for d as follows:

1. For each unique candidate word from all available contexts of d, add a vertex in G.
2. Add an undirected edge between two vertices v_i and v_j if the words corresponding to these vertices occur within a window of w contiguous tokens in any of the contexts.
3. The weight w_{ij} of an edge $(v_i, v_j) \in E$ is given as

$$w_{ij} = w_{ji} = \sum_{t \in T} \sum_{c \in C_t} \lambda_t \cdot \text{cossim}(c, d) \cdot \#_c(v_i, v_j)$$

We score vertices in G using their PageRank obtained by recursively computing:

$$s(v_i) = (1 - \alpha) + \alpha \sum_{v_j \in Adj(v_i)} \frac{w_{ji}}{\sum_{v_k \in Adj(v_j)} w_{jk}} s(v_j)$$

[Page et al.(1999)Page, Brin, Motwani, & Winograd].

Parameterized Edge Weights in CiteTextRank

- Unlike simple graph edges with fixed weights, our equations correspond to *parameterized* edge weights.
- We incorporate the notion of "importance" of contexts of a certain type using the λ_t parameters.

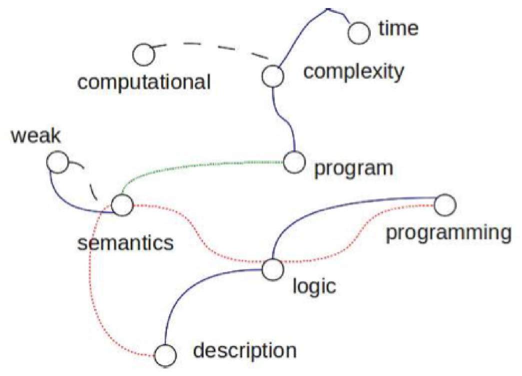


Figure : A small word graph. Edges from different contexts are shown using different colors/line-styles.

Datasets

Conference	#Titles(Org)	#Titles(CiteSeer)	#Queries	AvgKeywords	AvgCitingContexts	AvgCitedContexts
AAAI	5676	2424	93	4.15	9.77	13.95
UMD	490	439	163	3.93	20.15	34.65
WWW	2936	1350	425	4.81	15.91	17.39
KDD	1829	834	365	4.09	18.85	16.82

Results

How sensitive is CiteTextRank to its parameters?

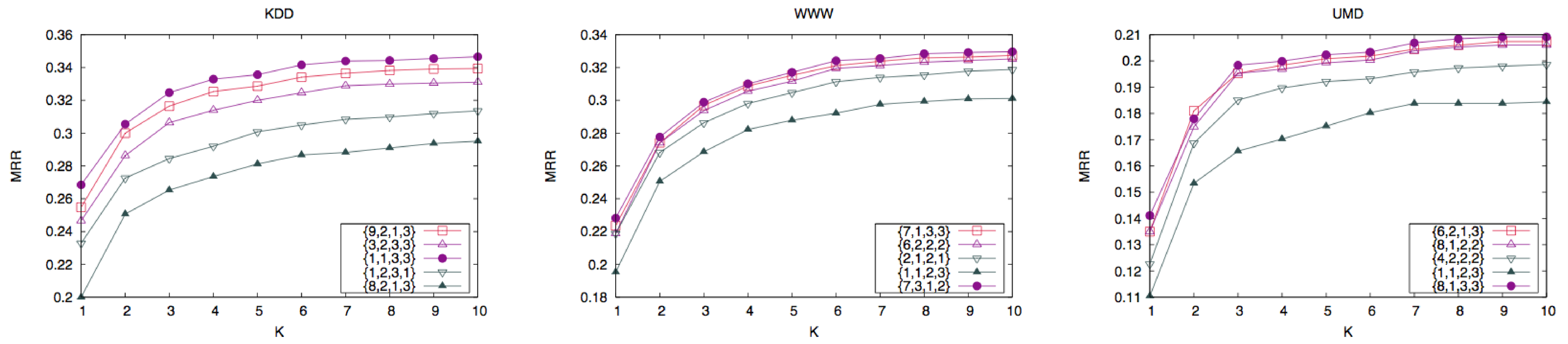


Figure : Parameter tuning for CTR. Sample configurations are shown. Setting a,b,c,d indicates window parameter is set to 'a' and the weights for content, cited and citing contexts set to 'b', 'c' and 'd', respectively.

How well does citation network information aid in key phrase extraction for research papers?

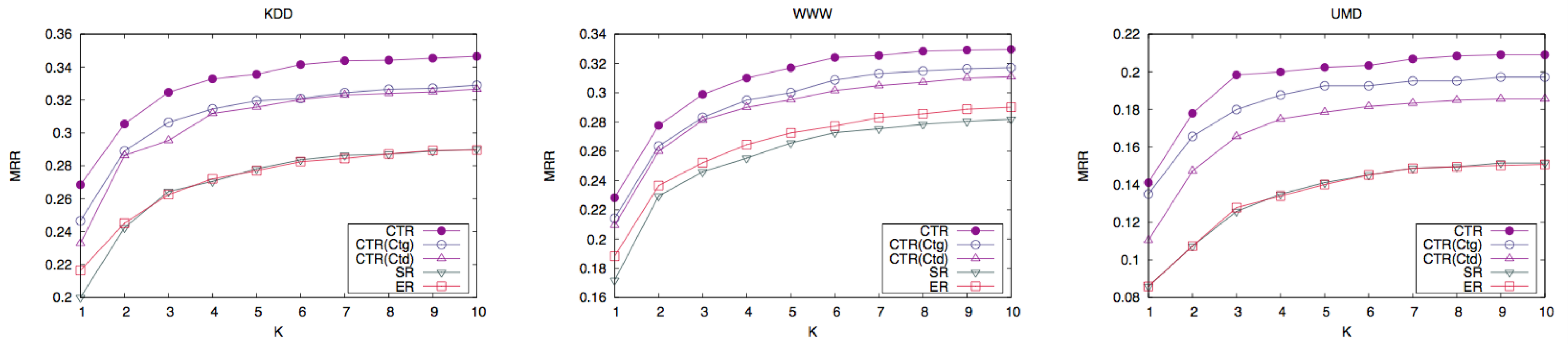


Figure : Effect of citation network information on keyphrase extraction. CTR that uses citation network neighbors is compared with ExpandRank (ER) that uses textually-similar neighbors and SingleRank (SR) that only uses the target document content.

How does CiteTextRank compare with other existing state-of-the-art methods?

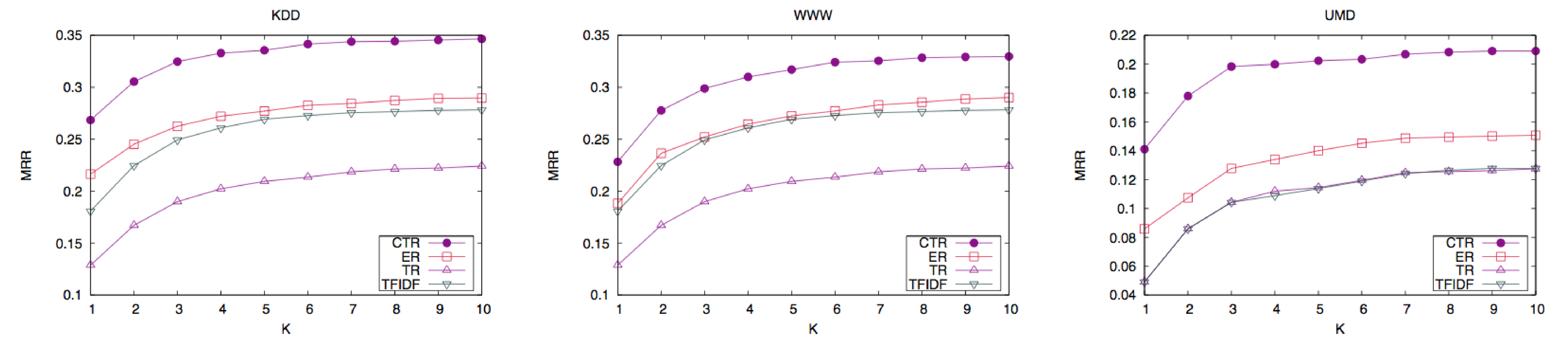


Figure : MRR curves for different keyphrase extraction methods. CiteTextRank (CTR) is compared with the baselines: TFIDF, TextRank (TR), and ExpandRank (ER).

Conclusions

- We proposed CiteTextRank (CTR), a flexible, unsupervised graph-based model for ranking keyphrases using multiple sources of evidence:
  - The textual content of a document and its citing and cited contexts in the interlinked document network.
- CTR gives significant improvements over baseline models for multiple datasets of research papers in the Computer Science domain.
- Future directions:
  - Further evaluation of CTR on other domains.
  - Extend CTR for extracting document summaries.

References

Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010).
Automatic keyphrase extraction via topic decomposition.
In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '10).

Mihalcea, R. & Tarau, P. (2004).
Textrank: Bringing order into text.
In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999).
The pagerank citation ranking: Bringing order to the web.
Technical report.

Shi, X., Leskovec, J., & McFarland, D. A. (2010).
Citing for high impact.
In Proceedings of the Joint Conference on Digital Libraries (JCDL '10).

Wan, X. & Xiao, J. (2008).
Single document keyphrase extraction using neighborhood knowledge.
In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI '08).

Acknowledgements: This research is partially supported by a grant from NSF to Cornelia Caragea. We thank Kishore Neppalli for helping with dataset construction tasks.

WWW: http://www.cse.unt.edu/~ccaragea